PREDICTING CUSTOMER CHURN USING MACHINE LEARNING

Authors

Lance Pollack, Riley Yee, Noah Chance, Waleed El-JAck

Affiliations

MAS648, Miami Herbert Business School

1 — INTRODUCTION

Customer churn directly impacts revenue and long-term business stability. Identifying customers who are likely to leave enables companies to take targeted, cost-effective retention actions.

This project develops a machine learning workflow to:

1. Understand what **churn prediction looks like** using an interpretable model.
2. Build a **high-performance predictive model** capable of capturing complex patterns.
3. Identify the **key drivers of churn** and translate them into actionable insights.

To make the prediction task intuitive, a **small Decision Tree** is first used to show how simple rules segment customers into higher- or lower-risk groups. We then apply **XGBoost** as the final model, leveraging its ability to capture nonlinear patterns and deliver stronger predictive performance.

2 — DATASET & EXPLORATORY ANALYSIS

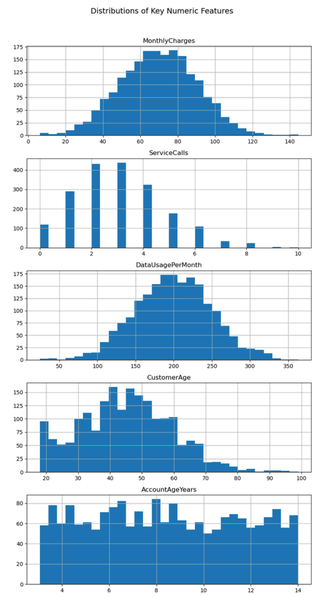
The dataset contains ~2000 customer records with a binary churn label (0 = No Churn, 1 = Churn).
Features span customer demographics, account characteristics, billing preferences, service interactions, and engineered behavioral indicators.

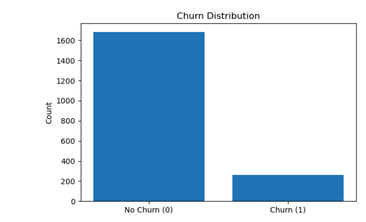
Key Dataset Characteristics

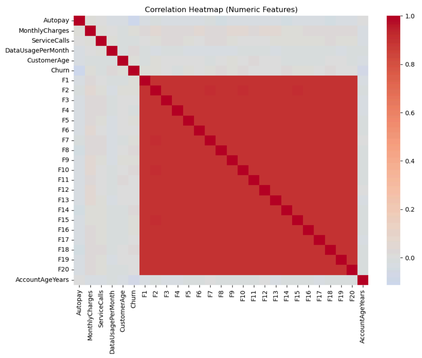
- Churn rate is highly imbalanced (~14%), raising a challenge for predictive modeling.
- Numerical features such as MonthlyCharges, DataUsagePerMonth, and AccountAgeYears (derived from DateOfServiceStart) show smooth, well-behaved distributions.
- ServiceCalls is right-skewed, indicating potential dissatisfaction signals.
- The engineered block of features (F1–F20) exhibits internal correlation patterns, suggesting latent structure in customer behavior.

Exploratory Analysis Highlights

- Customers who churn often differ in contract type, billing method, and usage patterns.
- Correlation heatmaps reveal clusters of related features and opportunities for tree-based models to exploit feature interactions.





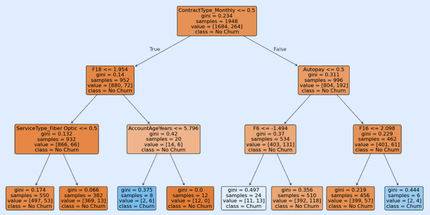


3 — INTERPRETABLE MODEL: DECISION TREE

To make the churn prediction task intuitive, a small Decision Tree model was used as an interpretable baseline.

The tree demonstrates how simple rules—such as contract type, autopay enrollment, service call frequency, and customer tenure—separate customers into higher- and lower-risk groups.

Although the tree is not the final predictive model, it clearly shows the structure of churn risk and provides a visual foundation for understanding how more advanced models, like XGBoost, build upon these patterns.



4 — XGBOOST MODELING PIPELINE

XGBoost was selected as the final model because it captures nonlinear relationships and feature interactions that simpler models miss.

Preprocessing

- Missing values handled using KNNImputer (k = 25)
- One-hot encoded categorical features
- Tree-based model → no scaling needed
- Evaluated with 10-fold stratified cross-validation using ROC–AUC

Two-Stage Hyperparameter Tuning

A multi-stage process was used to efficiently locate the best model:

1. RandomizedSearchCV:
   - Broad exploration across many depths, learning rates, subsampling levels, and regularization strengths to identify high-performing regions.
2. GridSearchCV:
   - A focused search around the most promising values for precise refinement.

Final Model Parameters

- n_estimators: 1500
- learning_rate: 0.15
- max_depth: 6
- min_child_weight: 20
- subsample: 0.9
- colsample_bytree: 0.8
- gamma: 10
- reg_alpha: 3.75
- reg_lambda: 0.85
- Imputation: KNNImputer (25 neighbors)

This configuration balanced depth, learning speed, and regularization, producing the strongest and most stable model.

5 — MODEL PERFORMANCE: XGBOOST

Using the optimized parameters, the final XGBoost model delivered strong and stable predictive performance.

Performance Metrics

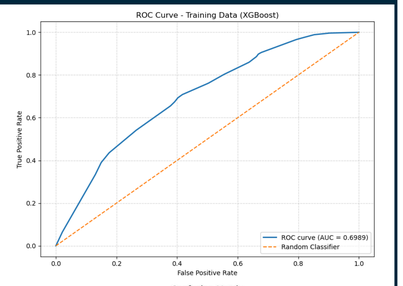
- Cross-validated AUC: 0.6933
- *Kaggle AUC Score: 0.6867*

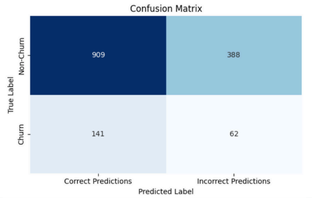
ROC Curve Interpretation

The ROC curve shows clear separation between churners and non-churners, demonstrating strong ability to rank customers by churn risk and performance meaningfully above chance.

Confusion Matrix Insights

- High true-negative accuracy
- Reasonable identification of churners despite imbalance
- Classification threshold can be adjusted depending on retention strategy priorities (e.g., capturing more churners vs. reducing false positives)





6 — FEATURE ANALYSIS

XGBoost provides ranked feature importances that reveal which customer attributes contribute most to predicting churn.

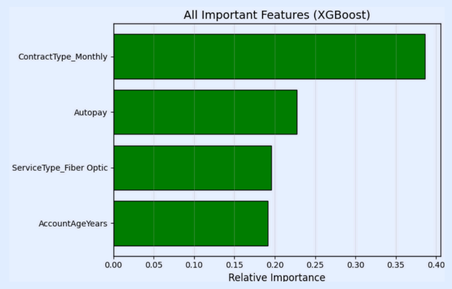
Key Predictors Identified by the Final Model

1. **ContractType_Monthly**: The strongest predictor of churn. Customers on month-to-month contracts consistently exhibit the highest churn risk due to flexibility and lack of long-term commitment.
2. **Autopay Enrollment**: A major stabilizing factor. Customers enrolled in autopay churn far less frequently, likely due to reduced billing friction and more consistent engagement.
3. **ServiceType_Fiber Optic**: Frequent support interactions strongly correlate with churn. This feature captures dissatisfaction or unresolved issues.
4. **AccountAgeYears**: Longer-tenure customers are less likely to churn, reflecting customer loyalty and established engagement patterns.

Contract type, payment behavior, and service setup are the primary levers driving churn.

Potential interventions:

- Convert monthly users to longer-term plans
- Promote autopay enrollment
- Address service-related concerns early



7 — BUSINESS TAKEAWAYS

Based on the XGBoost predictions and the risk patterns identified, the following actions are recommended:

1. Prioritize Month-to-Month Customers for Retention Campaigns

- These customers consistently represent the highest churn risk.
- Offering incentives—such as discounted 1-year plans or added benefits—could meaningfully reduce churn.

2. Encourage Autopay Enrollment

- Autopay is a strong stabilizing factor.
- A simple opt-in campaign (email, SMS, in-app prompt) could reduce churn at minimal cost.

3. Proactively Follow Up with Customers with Multiple Service Calls

- High call frequency signals dissatisfaction.
- Flagging these customers for early outreach may prevent churn before it occurs.

4. Strengthen Onboarding for Newer Customers

- Tenure strongly predicts loyalty.
- Improving onboarding support and first-month experience could boost long-term retention.

A churn-prevention strategy that combines risk scoring from the model with targeted interventions for high-risk segments offers the greatest impact. The XGBoost predictions provide a practical roadmap for prioritizing customers who benefit most from retention efforts.

8 — CONCLUSION

This project developed a complete machine learning workflow to understand and predict customer churn. A small Decision Tree provided an interpretable introduction, showing how simple rules segment customers into distinct risk groups. Building on that foundation, the optimized XGBoost model delivered the strongest predictive performance, successfully capturing nonlinear patterns and meaningful interactions among customers.

The results show that churn can be effectively predicted using a focused set of features, and that tree-based models are well-suited for identifying actionable retention signals. Overall, the workflow demonstrates how combining interpretability with advanced modeling offers both clarity and practical value, supporting data-driven decisions to reduce customer churn and improve long-term customer retention.